LE BIG DATA EST-IL L'ELDORADO DES GOUVERNEMENTS OUVERTS ?

par **Thomas BIZET**, Juriste à la CNIL, Doctorant en droit à l'Université Paris 1 Panthéon-Sorbonne (France).

À la fin du 17e siècle, des scientifiques et économistes anglais se mettent à espérer avec Sir William Petty une « arithmétique politique »[1]. Cette science est définie par Charles Davenant, un élève de Sir Petty comme « l'art de raisonner avec des chiffres sur des objets relatifs au gouvernement. »[2]

En 1751, Diderot inclut dans l'Encyclopédie un article relatif à cette arithmétique particulière, il y écrit que ces « opérations ont pour but des recherches utiles à l'art de gouverner les peuples, telles que celles du nombre des hommes qui habitent un pays ; de la quantité de nourriture qu'ils doivent consommer ; du travail qu'ils peuvent faire ; du temps qu'ils ont à vivre ; de la fertilité des terres ; de la fréquence des naufrages, etc. On conçoit aisément que ces découvertes et beaucoup d'autres de la même nature, étant acquises par des calculs fondés sur quelques expériences bien constatées, un ministre habile en tirerait une foule de conséquences pour la perfection de l'agriculture, pour le commerce tant intérieur qu'extérieur, pour les colonies, pour le cours et l'emploi de l'argent, etc. Mais souvent les ministres (je me garde de parler sans exception) croient n'avoir pas besoin de passer par des combinaisons et des suites d'opérations arithmétiques : plusieurs s'imaginent être doués d'un grand génie naturel, qui les dispense d'une marche si lente et si pénible, sans compter que la nature des affaires ne permet ni ne demande presque jamais la précision géométrique. Cependant si la nature des affaires la demandait et la permettait, je ne doute point qu'on ne parvînt à se convaincre que le monde politique, aussi bien que le monde physique, peut se régler à beaucoup d'égards par poids, nombre et mesure. »[3]

Aujourd'hui, la « nature des affaires » semble permettre, d'approcher tout du moins, la « précision géométrique ». L'évolution extraordinaire des technologies de traitement de données permet d'entrevoir des développements phénoménaux dans la tradition de « l'arithmétique politique ». Ces développements sont influencés notamment par la démarche de

[1] Sir W. PETTY, *Several Essays in Political Arithmetic,* 4e édition, Londres, 1960. Consulté le 20 décembre 2016 sur : https://archive.org/stream/severalessaysin00pettgoog.

[2] Cité par J. A. SCHUMPETER, *Histoire de l'analyse économique*, Vol. 1, Paris: Gallimard, 1983.

[3] D. DIDEROT, *Encyclopédie*, Volume III, 1751-1765. Consulté le 20 décembre 2016 sur : http://classiques.uqac.ca/classiques/Diderot_denis/encyclopedie/arithmetique_politique/arithmetique_pol.html.

recensement et d'ouverture des données détenues par les administrations dans le contexte du « Gouvernement Ouvert ».

Les chiffres sont en train de dévorer le monde pour paraphraser la citation de Marc Andreessen avec l'explosion de la création, de la collecte et de la circulation des informations et des capacités et des modalités de traitements de celles-ci.

Pour saisir cette explosion, le terme « Big data » a été utilisé de très nombreuses fois. Il sera ici utilisé pour faire référence aux grandes quantités d'informations recueillies sur de nombreuses personnes ou choses utilisant de nombreux périphériques[4] et aux traitements de ces informations. En effet, plus que la seule volumétrie, ce qui caractérise le "Big data" c'est la capacité à relier des données avec d'autres jeux de données, à les agréger et à chercher autant dans le contenu même de ces données que dans les informations contextuelles sur celles-ci[5]. Le volume et la variété des données permettent d'accroître la précision des algorithmes - les modèles utilisés pour traiter les données - par exemple pour effectuer des recherches dans le champ des analyses prédictives[6].

L'accroissement des puissances de calcul de ces technologies, allié aux nombreuses sources de données disponibles, peut faire naître l'ambition de collecter autant de données que possible sur toutes les sources possibles, de les analyser en temps réel et de prendre une décision optimale basée sur les circonstances actuelles plutôt que sur une projection idéalisée[7] : le rêve de l'algorithmique politique du 17e siècle.

Cette technologie est aujourd'hui assurément dans les mains et les systèmes d'information de nombreuses grandes entreprises. Toutefois, les États, par le biais notamment des « chief data officer », de l'ouverture de nombreux jeux de données et donc de leurs recensements, commencent très sérieusement à s'orienter vers ses nouvelles technologies pour affiner la prise de décision à défaut de l'automatiser. Ainsi, dans l'ouvrage collectif Beyond Transparency : Open Data and the Future of Civic Innovation, Tim O'Reilly[8] appelle par exemple à utiliser cette nouvelle

---

[4] P. N. HOWARD, S. SHOREY, S. C. WOOLLEY & M. GUO, *Creativity and Critique: Gap Analysis of Support for Critical Research on Big Data,* Oxford, UK: Project on Computational Prpaganda, 2016. Consulté le 3 décembre 2016, sur : https://papers.ssrn.com /sol3/papers.cfm?abstract_id=2822389.

[5] D. BOYD & K. CRAWFORD, "Critical Questions for Big Data", *Information, Communication & Society, 15*(5), 2012, pp. 662-679. Consulté le 3 décembre 2016, sur : https://people.cs.kuleuven.be/~bettina.berendt/teaching/ViennaDH15/boyd_crawford_2012.pdf.

[6] K. CRAWFORD & J. SCHULTZ, Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, *Boston College Law Review, 55*(1), 2014, pp. 93-128. Consulté le 3 décembre 2016, sur : https://papers.ssrn.com/sol3/ papers.cfm?abstract_id=2325784.

[7] E. MOROZOV, The Planning Machine : Project Cybersyn and the origins of the Big Data nation, *The New Yorker,* 13 octobre 2014. Consulté le 3 décembre 2016, sur : http://www.newyorker.com/magazine/2014/10/13/planning-machine

[8] Fondateur de O'Reilly Media, une maison d'édition spécialisée dans l'informatique et « futurologue ».

puissance de calcul pour développer des régulations directement par les algorithmes[9].
Il est aisé de pressentir que l'utilisation d'algorithmes par les administrations permettrait d'améliorer les services publics, d'en créer de nouveaux et d'en multiplier les usagers. De même, l'avènement de l'algorithme pourrait renforcer l'application des lois de Rolland. L'algorithme applique le même modèle pour tous – égalité – ne connaît ni le sommeil ni le droit de grève pour peu que les serveurs soient robustes – continuité – et il peut être modifié aisément, sa modification valant pour tous les calculs postérieurs – mutabilité.
À la suite du « Code is Law »[10] de Lawrence Lessig, il pourrait alors être possible d'imaginer un « Data is Government ».
Toutefois, ces évolutions doivent nécessairement amener à analyser les différents mécanismes qui sous-tendent une pratique « politique » des technologies « Big data ».
L'objectif de cette analyse est de présenter les défis actuels ou prospectifs d'une pratique balbutiante qui mérite une approche critique pour en éviter les abus ou les erreurs pouvant impacter directement les administrés « algorithmés ».

§ 1 – UN PREMIER DÉFI : D'OÙ PROVIENNENT LES DONNÉES ?

La pratique « Big data » suppose comme son nom l'indique de nombreuses données. Ces données dans le cadre d'une pratique tournée vers l'administration sont en grande partie des données collectées directement par les administrations, ou indirectement par des délégations de service public. Ces données sont pour la plupart collectées de longue date pour mesurer et gérer les actions publiques et constituer les statistiques publiques.
La majorité de ces données sont des données purement statistiques de sorte qu'elles ne permettent pas d'identifier, directement ou indirectement des individus.
Dans ce cadre, le défi pour une initiative « Big data » sera de faire communiquer les administrations détentrices de données afin de leur faire partager les données pour développer de nouvelles données et surtout de nouveaux services aux usagers, ou des économies d'échelles. Cette stratégie de « désîlotage » des systèmes d'information des administrations, dite stratégie « État Plateforme » est conduite par la direction interministérielle du numérique et système d'information et de communication de l'État (DINSIC) dirigée par Henri Verdier[11].

9 E. MOROZOV, The Planning Machine : Project Cybersyn and the origins of the Big Data nation, *The New Yorker,* 13 octobre 2014. Consulté le 3 décembre 2016, sur : http://www.newyorker.com/magazine/2014/10/13/planning-machine.
10 L. LESSIG, Code Is Law : On Liberty in Cyberspace, *Harvard Magazine*, 2000. Consulté le 20 décembre 2016 sur : http://harvardmagazine.com/2000/01/code-is-law-html
11 Voir notamment H. VERDIER & N. COLIN, *L'âge de la multitude : Entreprendre et gouverner après la révolution numérique,* 2012.

Un objectif soutenu par la DINSIC est donc de permettre à ces données d'être recenser et rendues interopérables afin de pouvoir être agrégés, partagés et réutilisés dans une démarche d'ouverture des données vers le secteur privé (de type « Open Data ») que vers le secteur public lui-même.
Si la plupart des données détenues par l'administration sont des statistiques, certaines données peuvent permettre d'identifier directement ou indirectement par recoupement des personnes.
La loi n°78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, sociales et fiscales, dite « Loi CADA », encadrait la réutilisation de ces données par le secteur privé dans un article 13 qui disposait que :

> « Les informations publiques comportant des données à caractère personnel peuvent faire l'objet d'une réutilisation soit lorsque la personne intéressée y a consenti, soit si l'autorité détentrice est en mesure de les rendre anonymes ou, à défaut d'anonymisation, si une disposition législative ou réglementaire le permet.
> La réutilisation d'informations publiques comportant des données à caractère personnel est subordonnée au respect des dispositions de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés ».

Cet article a été codifié au sein du Code des relations entre le public et l'administration (CRPA) par la loi n°2016-1321 du 7 octobre 2016, dite « Loi Lemaire », et le décret n°2016-308 du 17 mars 2016. L'article L322-2 du CRPA dispose que « la réutilisation d'informations publiques comportant des données à caractère personnel est subordonnée au respect des dispositions de la loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés » tandis que l'article R322-3 précise que « lorsque la réutilisation n'est possible qu'après anonymisation des données à caractère personnel, l'autorité détentrice y procède sous réserve que cette opération n'entraîne pas des efforts disproportionnés. »
L'article L321-1 du CRPA ouvre, dans ce cadre, la réutilisation de ces données « par toute personne qui le souhaite à d'autres fins que celles de la mission de service public pour les besoins de laquelle les documents ont été produits ou reçus**.** »
Toutefois, l'article L321-2 du CRPA exclu précisément du champ de la réutilisation « l'échange d'informations publiques entre les administrations, aux fins de l'exercice de leur mission de service public ». Dans le cas où ces échanges concerneraient des informations contenant des données à caractère personnel, les administrations devraient donc se conformer essentiellement à la loi n°78-17 modifiée. Cette conformité implique notamment des obligations d'information des personnes, d'exercice des droits des personnes concernées et plus globalement les traitements devraient avoir une finalité explicite, légitime et loyale. Ces obligations semblent complexes à mettre en place dans un

contexte « Big data » où la proportionnalité des données traitées n'est pas clairement définie, tout du moins initialement.
Si assez peu d'informations publiques contiennent actuellement des données à caractère personnel permettent d'identifier directement des individus, certaines données ne permettent cette identification qu'après de nombreux traitements. Les capacités de calcul et les traces que les individus laissent ne permettent que très difficilement de réaliser une anonymisation claire et efficace[12].
Toutefois, il convient de noter que dans le cas d'une démarche « Big data » de l'administration, cette démarche s'inscrira nécessairement dans l'exécution d'une mission de service public dont est investi le responsable ou le destinataire du traitement au sens de l'article 7 de la loi n°78-17 modifiée. Cette démarche ne devra donc pour autant pas faire oublier que « les poids, nombre et mesure » de Diderot sont souvent des informations sur des personnes[13] concernées par ces traitements[14], initialement ou finalement.

§ 2 – UN DEUXIEME DEFI : LA CONFIANCE DANS LES ALGORITHMES

Amasser les données n'est pas la finalité d'une démarche « Big data », l'objectif est de développer des modèles permettant d'améliorer une situation. Ces modèles, désignés souvent comme algorithmes, sont des suites d'opérations ou d'instructions permettant d'obtenir un résultat. Il s'agit du traitement des données en tant que tel.
Ces algorithmes sont le moteur d'un traitement « Big data ». Dans le cadre d'une démarche de « Gouvernement Ouvert », ce moteur doit être ouvert autant pour éviter une asymétrie d'informations générant des effets juridiques peu prévisibles que pour permettre à une communauté d'apporter une contribution à l'amélioration du modèle.
Par ailleurs, une ouverture permettrait à chacun de comprendre les modèles appliqués afin d'éviter la promesse du seul solutionnisme technologique[15], argument d'autorité d'un modèle proclamé vrai, car mathématique[16].

---

12 En ce sens : Y.-A. MONTJOYE, L. RADAELLI, V.K. SINGH & A. PENTLAND, Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science, 347*(6221), 30 janvier 2015, pp. 536-539. Consulté le 3 décembre 2016, sur : http://science. sciencemag.org/content/347/6221/536.

13 Voir en ce sens la disparition de la « personne » dans l'entretien d'Antoinette Rouvroy dans S. ABITEBOUL, C. FROIDEVAUX & A. ROUVROY, Big data : l'enjeu est moins la donnée personnelle que la disparition de la personne, *binaire*, 2016. Consulté le 3 décembre 2016, sur : http://binaire.blog.lemonde.fr/2016/01/22/le-sujet-de-droit-au-peril-de-la-gouvernementalite-algorithmique/.

14 Voir en ce sens l'approche « user-centric » préconisée par B. LEPRI, J. STAIANO, D. SANGOKOYA, E. LETOUZE & N. OLIVER, The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good, 2016. Dans B. LEPRI, J. STAIANO, D. SANGOKOYA, E. LETOUZÉ & N. OLIVER, *Transparent Data Mining for Big and Small Data,* Springer. Consulté le 3 décembre 2016, sur https://arxiv.org/abs/1612.00323

15 E. MOROZOV, *Pour tout résoudre cliquez ici – l'aberration du solutionnisme technologique* (trad. M-C. Braud), FYP, 2014.

16 E. MEDINA, Rethinking algorithmic regulation, *Kybernetes, 44*(6/7), 2015, pp. 1005-1019. Consulté le 3 décembre 2016, sur :

Mike Loukides, vice-président de la stratégie de contenu de O'Reilly Media, précise cette nécessité d'ouverture, en effet :

> « il n'y pas que les données qui doivent être ouvertes : il y a aussi les modèles ! […] Vous pouvez avoir toutes les données sur la criminalité que vous voulez, toutes les données de l'immobilier que vous voulez, toutes les données sur les performances des élèves que vous voulez, toutes les données médicales que vous voulez, mais si vous ne savez pas quels modèles sont utilisés pour générer des résultats, vous n'aurez pas beaucoup de réponses. »[17]

De la même manière, Cathy O'Neil, mathématicienne, précisait que :

> « vous ne savez pas vraiment ce que fait un modèle tant que vous ne pouvez pas interagir avec lui. Vous ne savez pas si un modèle est robuste tant que vous ne pouvez pas jouer avec ses paramètres. Enfin, vous ne savez pas si un modèle est le meilleur possible tant que vous n'avez pas laissé les gens essayer de l'améliorer. »[18]

En France, l'ouverture des modèles derrière des démarches « Big data » pourrait trouver racine dans l'article 15 de la Déclaration des droits de l'homme et du citoyen qui dispose que « la Société a le droit de demande compte à tout Agent public de son administration » incluant par extension les modalités de réalisation de cette administration. Plus précisément, la loi n°2004-801 du 6 août relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel, modifiant la loi n°78-17 du 6 janvier 1978, a créé un article 10 dans la loi n°78-17 modifiée disposant :

> « Aucune décision de justice impliquant une appréciation sur le comportement d'une personne ne peut avoir pour fondement un traitement automatisé de données à caractère personnel destiné à évaluer certains aspects de sa personnalité.
>
> Aucune autre décision produisant des effets juridiques à l'égard d'une personne ne peut être prise sur le seul fondement d'un traitement automatisé de données destiné à définir le profil de l'intéressé ou à évaluer certains aspects de sa personnalité.
>
> Ne sont pas regardées comme prises sur le seul fondement d'un traitement automatisé les décisions prises dans le cadre de la conclusion ou de l'exécution d'un contrat et pour lesquelles la personne concernée a été

http://wosc.co/wp-content/uploads/2016/03/Medina-Rethinking-Algorithmic-Regulation.pdf.

[17] M. LOUKIDES, We need open models, not just open data, *Radar*, 11 novembre 2014. Consulté le 3 décembre 2016, sur http://radar.oreilly.com/2014/11/we-need-open-models-not-just-open-data.html.

[18] C. O'NEIL, *Cool open-source models?,* 27 novembre 2013. Consulté le 3 décembre 2016, sur methbabe: https://mathbabe.org/2013/11/27/cool-open-source-models/.

> mise à même de présenter ses observations ni celles satisfaisant les demandes de la personne concernée. »

Poursuivant un objectif similaire la loi n°2016-1321 du 7 octobre 2016 a créé un article L311-3-1 du Code des relations entre le public et l'administration disposant que :

> « sous réserve de l'application du 2° de l'article L. 311-5, une décision individuelle prise sur le fondement d'un traitement algorithmique comporte une mention explicite en informant l'intéressé. Les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre sont communiquées par l'administration à l'intéressé s'il en fait la demande. »

En ce sens, dans le cas où la résolution d'un traitement algorithmique public produirait des effets juridiques à l'égard d'une personne, celle-ci devrait pouvoir présenter ses observations après communication des règles et des « principales caractéristiques » du traitement.

Enfin, la Commission d'accès aux documents administratifs (CADA) a été saisi pour obliger des administrations à communiquer des logiciels ou des codes sources de modèles. C'est le cas récemment du code source comprenant le modèle du portail « Admission post-bac (APB) » pour le traitement des candidatures post-baccalauréat sur les formations non sélectives[19]. Cette communication au public du code source peut être réalisée tant que le code source ne contient pas de données à caractère personnel – toutefois un code source ne devrait pas contenir en lui-même des données à caractère personnel hormis les noms de ses auteurs – et tant qu'il n'est pas protégé par des droits de propriété intellectuelle.

La loi exclut précisément la communication[20] et la réutilisation[21] des documents administratifs[22] – dont font partie les codes sources[23] – protégés par des droits de propriété littéraire et artistique. Cette protection, dans le cas des codes sources, peut empêcher l'administration de fournir tout ou partie du code

[19] Avis n°20161989 de la Commission d'accès aux documents administratifs.

[20] L'article L311-4 du Code des relations entre le public et l'administration dispose que « les documents administratifs sont communiqués ou publiés sous réserve des droits de propriété littéraire et artistique. »

[21] L'alinéa C de l'article L321-2 exclu des informations publiques réutilisables les informations contenues dans des documents sur « lesquels des tiers détiennent des droits de propriété intellectuelle ».

[22] Pour aller plus loin dans la distinction entre l'obligation de communication et les droits de réutilisations des informations publiques, voir W. GILLES, « Le Renouveau du droit à l'information à l'ère du numérique : entre obligation de publication de l'administration et affirmation du droit d'accès du citoyen ». *Revue Internationale de Droit des données et du Numérique*, 2016(2), 1-20. Consulté le 3 décembre 2016, sur : http://ojs.imodev.org/index.php/RIDDN/article/view/39.

[23] Dans son avis n°20144578 du 8 janvier 2015, la Commission d'accès aux documents administratifs « estime que les fichiers informatiques constituant le code source sollicité, produits par la direction générale des finances publiques dans le cadre de sa mission de service public, revêtent le caractère de documents administratifs, au sens de l'article 1er de la loi du 17 juillet 1978. »

source protégé[24]. L'ouverture des modèles nécessite donc parallèlement de repenser la stratégie « open source » des systèmes d'information des administrations[25].

L'ouverture des modèles est réalisée de manière encore moins enthousiaste que l'ouverture des seules données détenues par l'administration. Plus que les données, les modèles concentrent le pouvoir des sachants. Ces modèles peuvent apparaître comme objectivement neutres et justes, car mathématiques.

Or, cette neutralité des algorithmes est ardemment critiquée. Si effectivement les modèles, en tant qu'objet, sont neutres, leurs créateurs ne le sont peut-être pas[26].

Aujourd'hui, les modèles sont de plus en plus de nature automatisée, dans un processus technologique qui génère de la fascination, pourtant la discrétion humaine y joue un rôle toujours important.

Les analystes créant les modèles ont l'occasion ce faisant de laisser une empreinte idéologique – et potentiellement cachée – dans le processus[27]. Cette possibilité est ouverte depuis le début de la création du modèle. Les jeux de données doivent être activement construit, parfois en harmonisant ou rationalisant des jeux de données de sources différentes, cette activité nécessite diverses décisions (quelles bases de données utilisées, sur quel périmètre, etc.). D'autres décisions vont être plus subtiles, comme ce qui compte comme un « évènement » déclenchant telle ou telle opération du modèle, tout en éliminant les résultats qui pourraient être considérés comme faux[28]. Toutes ces étapes permettent d'exercer une discrétion humaine derrière le modèle.

Les modèles peuvent être biaisés par de multiples facteurs[29]. Ils peuvent l'être dans leur développement même[30] ou dans leur utilisation ultérieure[31].

---

[24] L'article L311-7 du Code des relations entre le public et l'administration précise que « lorsque la demande porte sur un document comportant des mentions qui ne sont pas communicables en application des articles L. 311-5 et L. 311-6 mais qu'il est possible d'occulter ou de disjoindre, le document est communiqué au demandeur après occultation ou disjonction de ces mentions. »

[25] Voir en ce sens X. BERNE, Comment l'Etat s'est ouvert à l'open source avec OpenFisca et Mes-aides, *NextInpact*, 2015. Consulté le décembre 20, 2016, sur : http://www.nextinpact.com/news/93605-comment-l-etat-s-est-ouvert-a-l-open-source-avec-openfisca-et-mes-aides.htm.

[26] J. C. MCGINTY, Algorithms Aren't Biased, But the People Who Write Them May Be. *The Wall Street Journal, 2016*. Consulté le 20 décembre 2016, sur : http://www.wsj.com/articles/algorithms-arent-biased-but-the-people-who-write-them-may-be-1476466555.

[27] K. A. BAMBERGER, Technologies of Compliance: Risk and Regulation in a Digital Age. *Texas Law Review, 88*(4), 2010, pp. 669-740. Consulté le 3 décembre 2016, sur : http://scholarship.law.berkeley.edu/facpubs/1665/.

[28] Voir en ce sens le « rapport minoritaire » de P. K. Dick.

[29] N. BYRNES, Why We Should Expect Algorithms to Be Biased, *MIT Technology Review*, 2016. Consulté le décembre 20, 2016, sur https://www.technologyreview.com/s/601775/why-we-should-expect-algorithms-to-be-biased/

[30] T. Z. ZARSKY, Transparent Predictions. *Illinois Law Review*, 27 août 2013, pp. 1503-1570. Consulté le 3 décembre 2016, sur https://www.illinoislawreview.org/wp-content/ilr-content/articles/2013/4/Zarsky.pdf.

Ces biais peuvent avoir des impacts conséquents. Par exemple, aux États-Unis d'Amérique, un modèle de prédiction des scores de récidives a été audité par le journal d'investigation ProPublica. Les journalistes ont découvert des disparités du modèle basées sur la couleur de peau des personnes concernées[32].

Pour autant, l'ampleur des biais est difficilement mesurable, en particulier dans le cas des algorithmes dits prédictifs. Cette catégorie d'algorithme peut souffrir de prédictions autoréalisatrices en s'autovalidant au fur et à mesure des prédictions et de leurs « réalisations ». Il est possible de retrouver par exemple ce biais dans le logiciel de « prédiction policière » intitulé « PredPol », un travail critique de Ismaël Benslimane présente les biais de ce logiciel[33].

Cachés derrière la neutralité mathématique, les modèles entraînent ainsi avec eux les biais de leurs créateurs. Les modèles doivent ainsi être ouverts pour vérifier leur fonctionnement.

Cette ouverture suppose une capacité ultérieure à auditer et comprendre le modèle. La communication du code source dans un format papier ne saurait pas permettre cet audit[34]. Les modèles doivent être examinés, et parfois contrôler, ex ante, afin de valider leur légitimité, et ex post, afin de valider leur utilisation[35].

La création d'une agence ou d'une autorité ad hoc est un sujet étudié par le Conseil National du Numérique en ce que concerne la « loyauté des plateformes ». À l'issue d'un rapport remis le 13 juin 2014 au ministre de l'Économie, du Redressement productif et du Numérique intitulé Neutralité des plateformes : Réunir les conditions d'un environnement ouvert et soutenable, le Conseil National du Numérique proposait notamment de garantir la loyauté du système des données. Le terme « neutralité » ayant été particulièrement dénoncé[36], le terme loyauté a été conservé. À la suite d'un rapport du Conseil Général de l'Économie[37] transmis le 13

---

[31] E. BOZDAG, "Bias in algorithmic filtering and personalization", *Ethics and Information Technology, 15*(3), septembre 2013, pp. 209-227. Consulté le 3 décembre 2016, sur : http://dl.acm.org/citation.cfm?id=2560640.

[32] J. ANGWIN, J. LARSON, S. MATTU & L. KIRCHNER, "Machine Bias", *ProPublica*, 23 mai 2016. Consulté le 3 décembre 2016, sur https://www.propublica.org /article/ machine-bias-risk-assessments-in-criminal-sentencing.

[33] I. BENSLIMANE, « Predpol : prédire des crimes ou des banalités ? », *Cortecs*, 10 décembre 2014. Consulté le 3 décembre 2016, sur https://cortecs.org/ mathematiques/ predpol-predire-des-crimes-ou-des-banalites/.

[34] E. BROUZE, Admission post-bac : « Le code est quasiment inexploitable » *Rue89,* 19 octobre 2016. Consulté le 3 décembre 2016, sur : http://rue89.nouvelobs.com/2016/10/19/admission-post-bac-code-est-quasiment-inexploitable-265455.

[35] En ce sens, D. K. CITRON, « Technological Due Process », *Washington University Law Review, 85*, 1249-1313, 2007. Consulté le 3 décembre 2016, sur : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1012360.

[36] Voir en ce sens les propositions de l'association La Quadrature du Net relatives à la loyauté des plateformes. Consulté le 21 décembre 2016 sur : https://www.laquadrature.net/files/propositions%20LQDN%20Loyaut%C3%A9%20des%20plateformes.pdf.

[37] I. PAVEL & J. SERRIS, *Modalités de régulation des algorithmes de traitement des contenus.* Conseil Général de l'Economie, 2016. Consulté le 20 décembre 2016, sur : http://www.economie.gouv.fr/files/files/directions_services/cge/Rapports/2016_05_13_Rapport_Algorithmes(1).pdf.

mai 2016 à la Secrétaire d'État chargée du numérique, l'INRIA et le Conseil National du Numérique sont chargés de prototyper et d'effectuer des expérimentations visant à « noter » les loyautés des grandes plateformes privées (GAFAM). Cette expérimentation – ni les précédents rapports – n'inclut pas les plateformes publiques et la loyauté des codes sources et algorithmes utilisés par les administrations dans le cadre de leurs missions de service public.

Ces audits peuvent également être réalisés dès le développement du modèle par des études de risque lors de sa création. Ces études de risques existent déjà en ce qui concerne la sécurité des systèmes d'information et vont se développer sur les risques « vie privée » avec le règlement européen applicable en 2018 créant les « études d'impact sur la vie privée ». Sans nécessiter un coût trop important les équipes de « data scientist » de la mission Etalab effectuent des accompagnements[38] et sur des modèles utilisant des données à caractère personnel la Cnil pourrait de même effectuer des accompagnements autant dans la définition des modèles que dans la conformité des traitements.

L'ouverture suppose la possibilité du public – et de la « multitude » - de s'emparer de même du sujet. Par ailleurs, comme le précise Mark Fenster, « généralement, le niveau d'expertise, de temps et d'attention disponibles en dehors des agences gouvernementales est plus important que la connaissance disponible dedans. »[39]

Ouvert, le code est contrôlable par des communautés d'experts « data scientist » qui peuvent en saisir les subtilités et en traduire le mécanisme précis à d'autres experts juridiques, sociologues, économistes, etc. C'est par exemple le cas du code « ouvert » du portail « APB » qui a été fourni dans un format non réutilisable et dont les variables sous-tendant le modèle n'ont pas été clairement commentées. Une communauté a traduit le code, l'a rendu lisible et compréhensible pour en extraire le modèle appliqué[40].

Toutefois, un dernier défi dans la confiance dans le modèle s'applique précisément au cas des technologies « Big data ». De nombreux modèles prédictifs ne sont pas des modèles fixés dont les « évènements » sont inamovibles. Pour améliore la pertinence de la prédiction, le modèle est développé pour s'améliorer au fur et à mesure de ses prédictions, il est « autoapprenant ». Cet apprentissage peut être supervisé, semi-supervisé ou non supervisé suivant la liberté laissée à l'algorithme. Ces méthodes ne permettent pas de contrôler *ex ante* ou *ex post* puisque la structure

[38] Voir en ce sens le très bon exemple d'ouverture de l'algorithme « Bob Emploi », consulté le 20 décembre 2016 sur https://agd.data.gouv.fr/2016/11/14/760/

[39] M. FENSTER, "The Opacity of Transparency", *Iowa Law Review*(91), 885-949, 2006. Consulté le 3 décembre 2016, sur : https://papers.ssrn.com/sol3/papers.cfm? abstract_id=928550

[40] S. GRAVELEAU, APB : les questions que soulève le code source, *Le Monde, 2016*. Consulté le 20 décembre 2016, sur : http://www.lemonde.fr/campus/article/2016/10/25/apb-les-questions-que-souleve-le-code-source_5020076_4401467.html.

même du modèle se modifie avec les résultats[41]. « Avec les Big Data, cependant, cette traçabilité va devenir beaucoup plus difficile. La base de prédiction d'un algorithme peut devenir beaucoup trop complexe pour qu'un être humain moyen la comprenne. »[42]

Dans ces cas précis, qui soulignons-le sont souvent les modèles présentés ou fantasmés derrière la terminologie « Big data », la transparence est une course derrière l'évolution du système de données[43].

Enfin, soulignons que la confiance dans les données traitées et les modèles utilisés supposent que ces données et ces modèles n'ont pas été modifiés frauduleusement pour en arriver au résultat. Cette sécurité est d'autant plus complexe à atteindre que l'ouverture des modèles entraîne la compréhension des mécanismes et rend plus aisées les modalités de manipulation de données en entrée.

§ 3 – UN TROISIÈME DÉFI : CONTINUER À CHERCHER LES CAUSES

Les technologies « Big data » invitent à s'intéresser aux similarités, aux corrélations[44], pour prédire des résultats et corriger des conséquences. Apporter des solutions aux problèmes, des produits aux frictions[45]. Si cette approche peut apporter des solutions concrètes à des conséquences constatées, elle n'induit pas l'identification de la cause des frictions.

Michael Flowers, l'ancien chef du bureau des statistiques de la ville de New York expliquait dans un entretien aux auteurs du livre *Big Data : À Revolution That Will Transform How We Live, Work and Think* qu'il n'était pas intéressé par les causes, « la causalité est pour les autres, et franchement c'est très risqué quand vous commencez à parler de la causalité… Vous savez, nous avons de véritables problèmes à résoudre. »[46] Les auteurs poursuivent en ce que « nous entrons dans un monde de prédictions basées sur des constantes qui pourraient ne pas être en mesure d'expliquer les raisons de nos décisions. »[47]

---

41 Voir notamment J. BURREL, "How the machine 'thinks': Understanding opacity in machine learning algorithms", *Big Data & Society, 3*(1), 10, 2016. Consulté le 3 décembre 2016, sur https://ssrn.com/abstract=2660674.

42 K. CUKIER & V. MAYER-SCHÖNBERGER, *Big Data: A Revolution That Will Transform How We Live, Work, and Think,* Eamon Dolan/Houghton Mifflin Harcourt, 2013.

43 T. Z. ZARSKY, "Transparent Predictions", *Illinois Law Review*, 27 aout 213, pp. 1503-1570. Consulté le 3 décembre 2016, sur : https://www.illinoislawreview.org/wp-content/ilr-content/articles/2013/4/Zarsky.pdf

44 T. VIGEN (s.d.), *Spurious correlations*. Consulté le 3 décembre 2016, sur : http://www.tylervigen.com/spurious-correlations.

45 Voir en ce sens le manifeste des Startups d'Etat : https://beta.gouv.fr/startups.html

46 Propos rapportés dans E. MOROZOV, "The Planning Machine : Project Cybersynand the origins of the Big Data nation", *The New Yorker*, 13 octobre 2014. Consulté le décembre 03, 2016, sur http://www.newyorker.com/magazine/2014/10/13/planning-machine.

47 K. CUKIER & V. MAYER-SCHÖNBERGER, *Big Data: A Revolution That Will Transform How We Live, Work, and Think,* Eamon Dolan/Houghton Mifflin Harcourt, 2013.

Cette poursuite de la statistique, d'un « data-driven model » disruptif cher à l'esprit startup appliqué aux politiques publiques ne doit pas pour autant faire oublier que le traitement des conséquences est un chiffre qui s'autoalimente au fur et à mesure des conséquences. Le traitement de la cause, plus subtil en termes de chiffres, s'efface en même temps que la cause.

Or, comme le décrit Cathy O'Neil, « une formule peut être parfaitement inoffensive en théorie. Mais lorsqu'elle est employée à grande échelle et devient un standard national ou mondial, elle crée sa propre économie déformée et dystopique.»[48] Dans un futur imaginaire régi par des décisions automatisées, celles-ci risqueraient de devenir des standards sans qu'il ne soit possible d'en expliquer la raison. Cette « gouvernementalité algorithmique »[49] balbutiante doit être accompagnée et faire l'objet d'un débat critique.

[48] C. O'NEIL *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,* Crown.

[49] A. ROUVROY & T. BERNS, « Gouvernementalité algorithmique et perspectives d'émancipation », *Réseaux*, 1(177), 163-196, 2013. Consulté le 3 décembre 2016, sur : http://www.cairn.info/resume.php?ID_ARTICLE=RES_177_0163.